Biased estimates in Mendelian randomization studies conducted in unrepresentative samples

Large cohort studies have transformed our ability to evaluate complex exposure-outcome relationships. However, these studies are typically not representative of the source population from which they are sampled, either due to selection at recruitment, or attrition over time (in prospective studies), or both. In UK Biobank, for example, the response rate was approximately 5%, and the resulting sample healthier and more highly educated than the general population of the UK (1).

It is often assumed that while this lack of representativeness is problematic for estimating prevalence, it will make little or no difference to association analyses. In our view this is overly optimistic - we have recently examined the potential impact of selection bias on results obtained from studies with low response rates (2). We argue that, because selection can induce collider bias (which occurs when two variables independently influence a third variable, and that variable is conditioned upon), selection can lead to biased estimates of associations.

In this context, it is worth considering the analysis in *JAMA Cardiology* by Lyall et al. (3), who reported an association of body mass index with cardiometabolic disease in UK Biobank. The exposure and several of the outcomes (e.g., coronary heart disease, type 2 diabetes) might plausibly be expected to be negatively associated with participation in UK Biobank, in which case a spurious negative association between the two will be created, or any true association biased, because the use of a sample subject to this selection amounts to conditioning on a collider.

Lyall et al. use Mendelian randomisation (4), an increasingly popular method to support stronger causal inference in observational data, but this does not protect against the potential bias we describe. In the Avon Longitudinal Study of Parents and Children (ALSPAC), a large prospective cohort study in the UK, both BMI and smoking predict attendance at subsequent assessment clinics (2). Genetic antecedents of factors such as BMI and smoking that influence participation will be subject to a similar bias, which we see in ALSPAC (unpublished data).

The extent to which bias due to selection will occur will depend on the particular association being explored, and the selection mechanisms operating. We suggest that researchers consider the potential for selection bias to be affecting their analyses, and carry out sensitivity analyses to assess robustness of their conclusions to selection bias.

Marcus Munafò PhD
George Davey Smith DSc

Author Affiliations: MRC Integrative Epidemiology Unit at the University of Bristol, Bristol, UK (Munafò, Davey Smith); UK Centre for Tobacco and Alcohol Studies, School of Experimental Psychology, University of Bristol, Bristol, UK (Munafò); School of Social and Community Medicine, University of Bristol, Bristol, UK (Davey Smith).

Corresponding Author: Marcus Munafò, PhD. School of Experimental Psychology, University of Bristol, 12a Priory Road, Bristol, BS8 1TU, United Kingdom (marcus.munafo@bristol.ac.uk).

1. Fry A, Littlejohns TJ, Sudlow C, Doherty N, Adamska L, Sprosen T, Collins R, Allen NE. Comparison of Sociodemographic and Health-Related Characteristics of

UK Biobank Participants with the General Population. Am J Epidemiol. 2017.

2. Munafò MR, Tilling K, Taylor AE, Evans DM, Davey Smith G. Collider Scope: When selection bias can substantially influence observed associations. Int J Epidemiol 2017.

3. Lyall DM, Celis-Morales C, Ward J, Iliodromiti S, Anderson JJ, Gill JMR, Smith DJ, Ntuk UE, Mackay DF, Holmes MV, Sattar N, Pell JP. Association of Body Mass Index With Cardiometabolic Disease in the UK Biobank: A Mendelian Randomization Study. JAMA Cardiol. 2017

4. Davey Smith D, Ebrahim S.'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? Int J Epidemiol. 2003;32(1):1-22.